Small coresets via negative dependence: DPPs, linear statistics, and concentration

Rémi Bardenet[1], Subhroshekhar Ghosh[2], Hugo Simon-Onfroy[3], Hoang Son Tran[2]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, [2]National University of Singapore [3]Université Paris-Saclay, CEA, Irfu

Overview

- A **coreset** is a subset of a (large) training set, such that minimizing the coreset empirical loss is a **controlled replacement for intractable minimization** of the global empirical loss.
- State-of-the-art coreset constructions rely on specific i.i.d. sampling, but recent works (Tremblay et al., 2019) provided empirical support for **Determinantal Point Processes** (DPPs), a **tunable and tractable form of negative dependence sampling**.
- We prove that DPP sampling outperforms state-of-the-art methods, i.e. **DPP sampling produces smaller coresets**. We do so by obtaining **concentration inequalities for general kernels** that extend well beyond coreset problem. Finally, we **empirically verify performances** on both synthetic and real data.

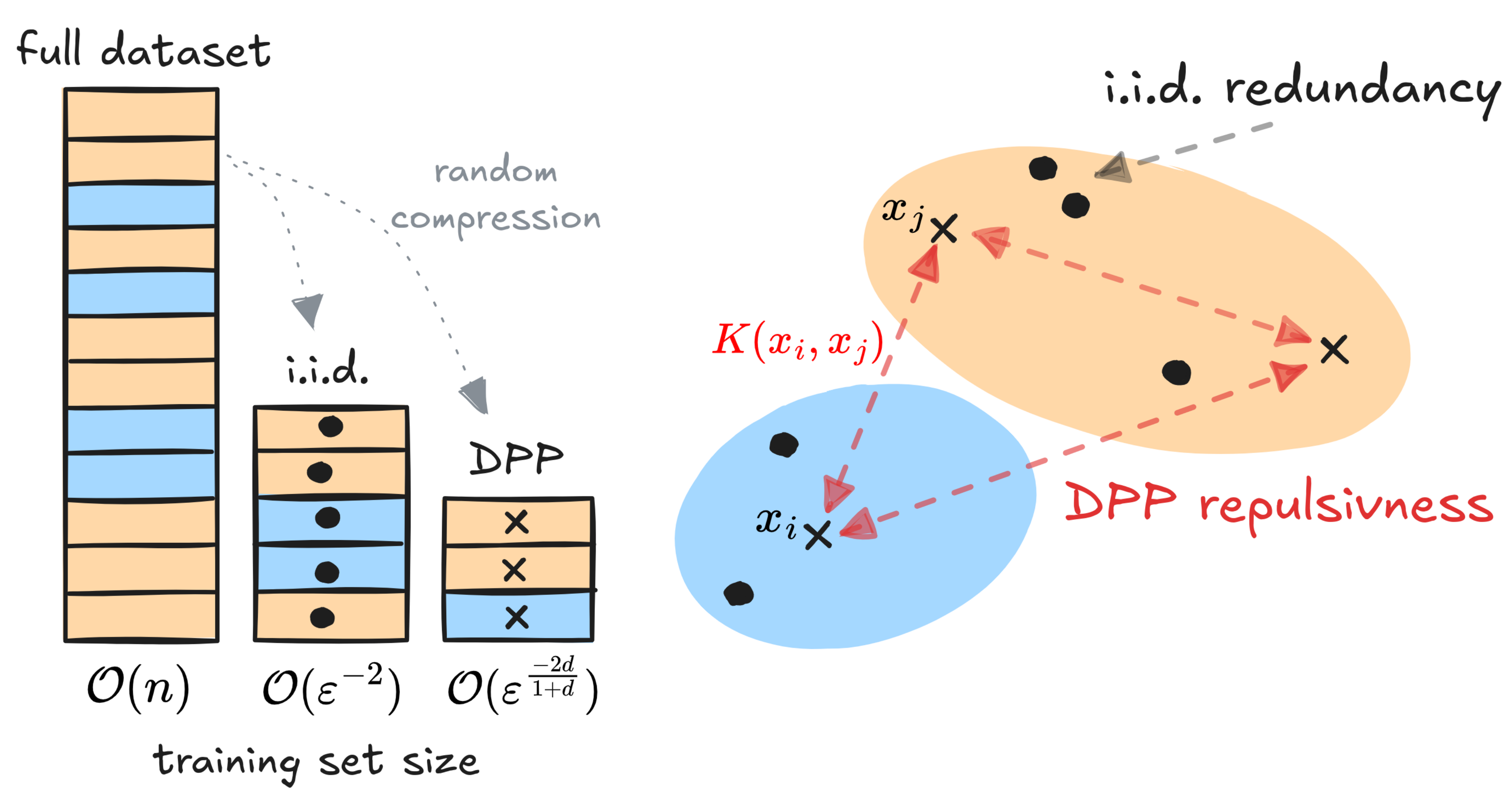


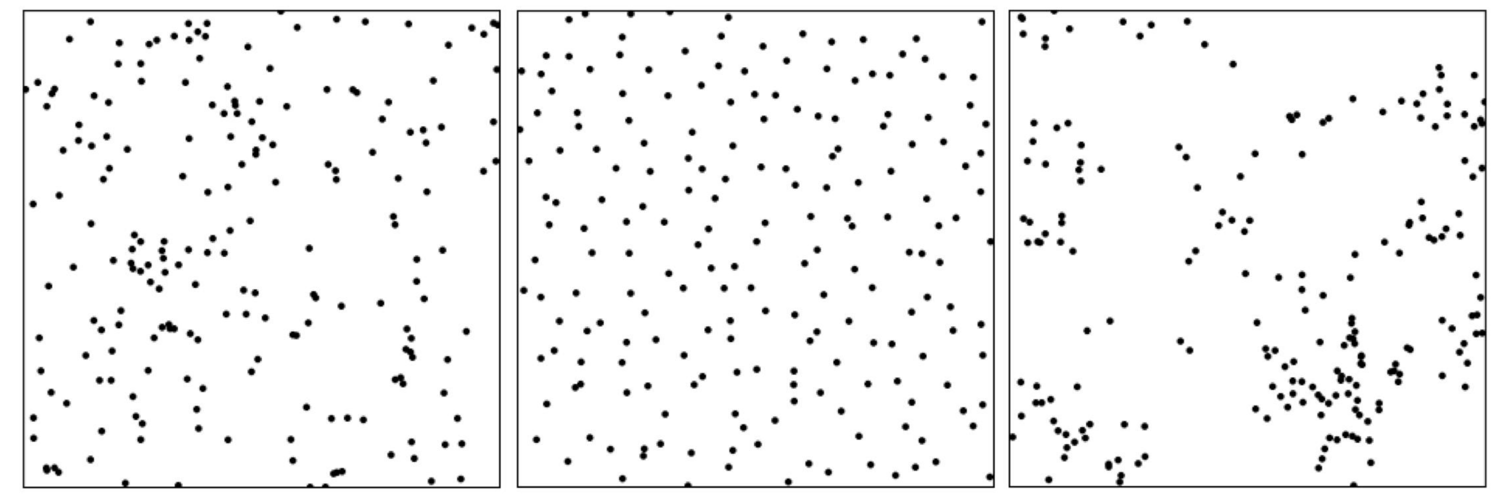
Figure 1: Negative dependence builds more representative subsets

Coresets

- Many ML problems can be stated as finding a query $f^* \in \mathcal{F}$ that **minimizes empirical loss** on dataset $\mathcal{X}$, $L(f) := \sum_{x \in \mathcal{X}} f(x)$
- However, optimization complexity grows with cardinality of dataset $n = |\mathcal{X}|$. What if we could perform **optimization on a compressed dataset** $\mathcal{S}$, of size m independent of n, and **still guarantee global optimization?** This idea is formalized by coresets.
- An ε**-coreset** is a subset $\mathcal{S} \subseteq \mathcal{X}$, possibly with weights ω, such that $L_\mathcal{S}(f) := \sum_{x \in \mathcal{S}} \omega(x) f(x)$ is within ε of $L(f)$, uniformly in $f \in \mathcal{F}$.
- ε-coreset constructions currently rely on i.i.d. **sensitivity sampling** (Bachem et al., 2017) with error rate $\varepsilon = \mathcal{O}(m^{-\frac{1}{2}})$, i.e. $m = \mathcal{O}(\varepsilon^{-2})$.

Determinantal Point Processes (DPPs)

Figure 2: Left: Poisson Center: Determinantal Right: Permanental



- A DPP is a random subset of points, whose correlation functions ρ are given by determinants of some kernel K. For any $n \in \mathbb{N}$,

$$\rho((x_i)_{i=1}^n) = \det(K(x_i, x_j))_{i,j=1}^n.$$

- Given a DPP $\mathcal{S}$ and a query function f, an unbiased estimator of global loss $L(f)$ is $L_\mathcal{S}(f) = \sum_{x \in \mathcal{S}} f(x)/K(x,x)$.

Theoretical results

- Let $\mathcal{S}$ be a (possibly non-symmetric) DPP over $\mathcal{X}$, and let $\mathcal{F}$ be a class of real-valued, bounded functions.

Theorem 1. *$\exists A$ abs. constant, s.t. $\forall f \in \mathcal{F}$, $\forall \varepsilon > 0$ small enough:*

$$\mathbb{P}\left(\left|\frac{L_\mathcal{S}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{\varepsilon^2}{4A\,\mathbb{V}\mathrm{ar}[L_\mathcal{S}(f)/L(f)]}\right).$$

- Application: Bardenet et al., 2021 constructed a DPP $\mathcal{S}$ termed **discretized multivariate OPE** which yields $\mathbb{V}\mathrm{ar}[L_\mathcal{S}(f)/L(f)] = \mathcal{O}(m^{-1-1/d})$, vs. $\mathcal{O}(m^{-1})$ for i.i.d. sampling.

Theorem 2. *Two common scenarios for space of queries $\mathcal{F}$:*

1. *If $\dim \mathrm{span}_\mathbb{R}(\mathcal{F}) = D < \infty$ (e.g. finite dim regression), then*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \left|\frac{L_\mathcal{S}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq 2 \exp\left(6D - C\varepsilon^2 m^{1+1/d}\right).$$

2. *If $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^D\}$, f_θ Lipschitz in θ (e.g. k-means), then*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \left|\frac{L_\mathcal{S}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq 2 \exp\left(C'D - D \log \varepsilon - C\varepsilon^2 m^{1+1/d}\right).$$

- This implies DPPs can build ε-coresets of size $m = \mathcal{O}(\varepsilon^{\frac{-2d}{1+d}}) \lesssim \mathcal{O}(\varepsilon^{-2})$

Experiments

Figure 3: Synthetic 2D trimodal dataset

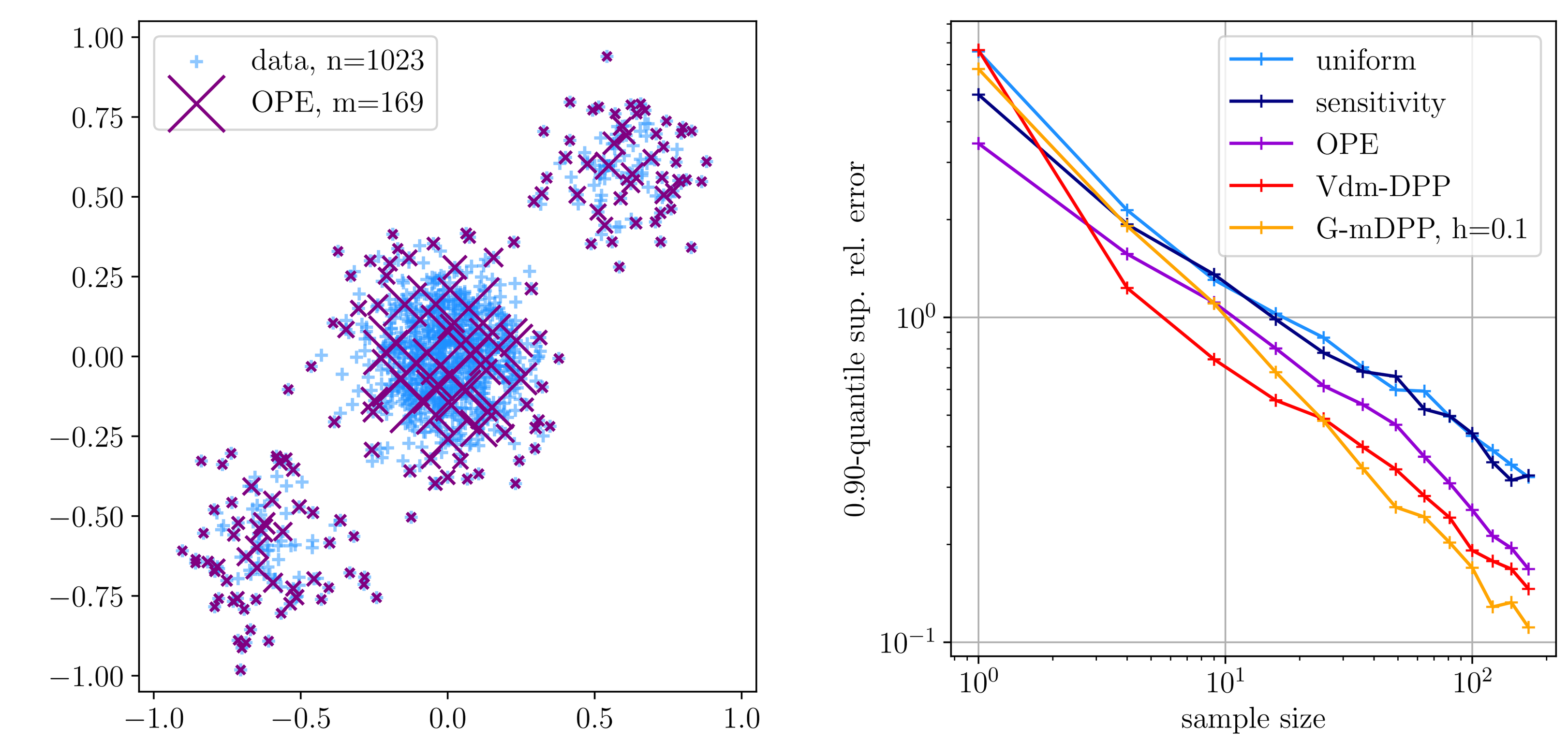
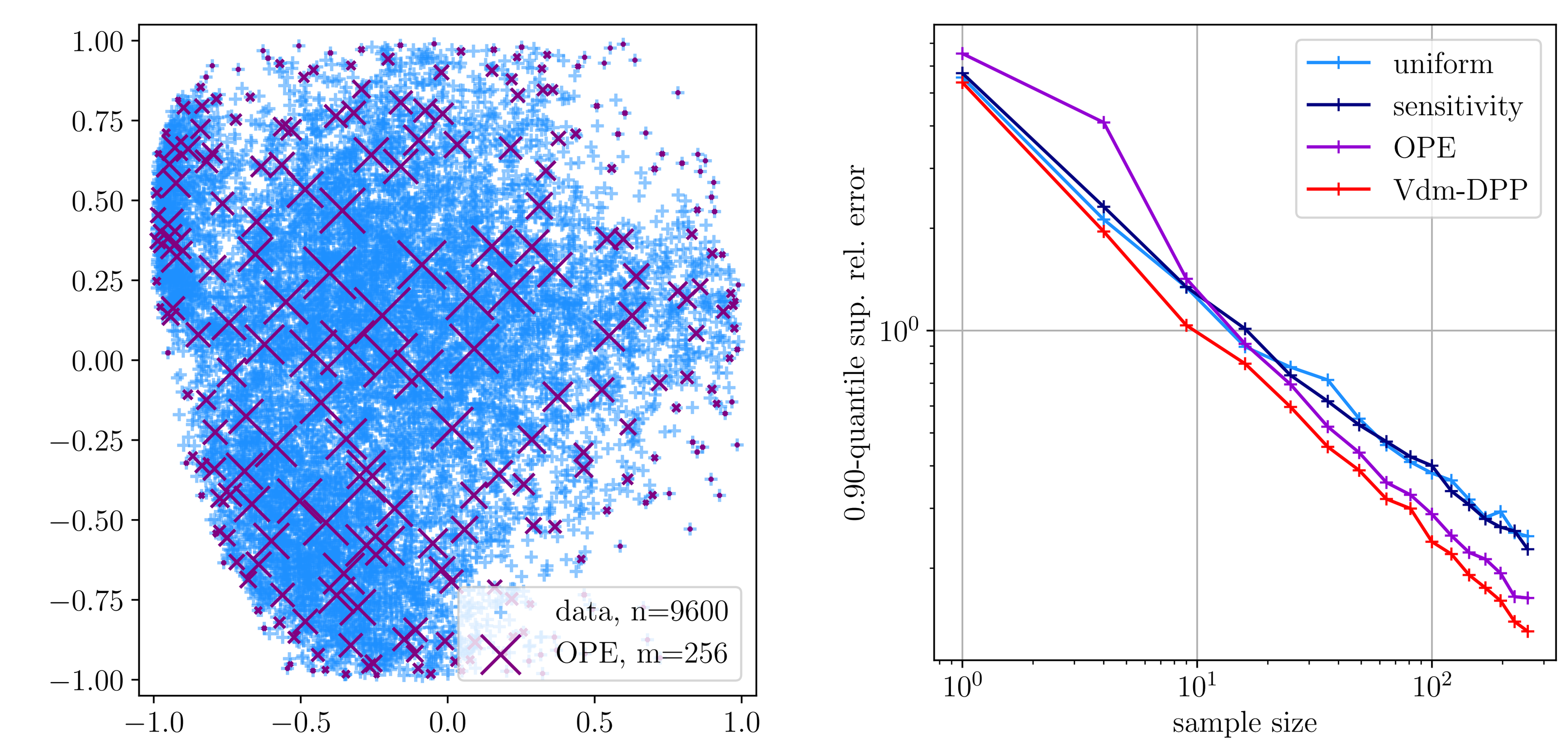


Figure 4: MNIST dataset (2D projection shown)



- Comparison of coreset samplers for k-means problem, on synthetic and MNIST dataset. We measure the 90%-quantile of the worst case relative error $\sup_f |L_\mathcal{S}(f)/L(f) - 1|$, depending on coreset size m.
- Measured error is about $\mathcal{O}(m^{-\frac{1}{2}-\frac{1}{2d}})$ for DPP samplers, while being about $\mathcal{O}(m^{-\frac{1}{2}})$ for i.i.d. samplers, consistent with theory.

References

Bachem, O., Lucic, M., & Krause, A. (2017). Practical coreset constructions for machine learning. *arXiv: Machine Learning*. https://api.semanticscholar.org/CorpusID:88517375

Bardenet, R., Ghosh, S., & Lin, M. (2021). Determinantal point processes based on orthogonal polynomials for sampling minibatches in sgd. *Advances in Neural Information Processing Systems*, *34*, 16226–16237.

Tremblay, N., Barthelmé, S., & Amblard, P.-O. (2019). Determinantal point processes for coresets. *Journal of Machine Learning Research*, *20*(168), 1–70.